

02.7.2004

日 本 国 特 許 庁
JAPAN PATENT OFFICE

REC'D: 26 AUG 2004

WIPO

PCT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: 2003年 7月 1日

出 願 番 号
Application Number: 特願2003-189716

[ST. 10/C]: [JP2003-189716]

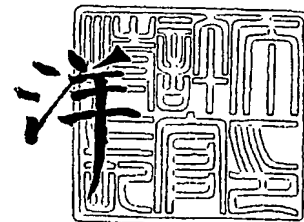
出 願 人
Applicant(s): 株式会社山武

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

2004年 8月12日

特許庁長官
Commissioner,
Japan Patent Office

小 川



出証番号 出証特2004-3071975

【書類名】 特許願

【整理番号】 20030113

【提出日】 平成15年 7月 1日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/20

【発明者】

 【住所又は居所】 東京都渋谷区渋谷 2 丁目 1 2 番 1 9 号 株式会社 山武
 内

 【氏名】 村上 英治

【発明者】

 【住所又は居所】 東京都杉並区和泉 3 丁目 1 5 番 1 6 号

 【氏名】 寺野 隆雄

【特許出願人】

 【識別番号】 000006666

 【氏名又は名称】 株式会社 山武

【代理人】

 【識別番号】 100064621

 【弁理士】

 【氏名又は名称】 山川 政樹

 【電話番号】 03-3580-0961

【手数料の表示】

 【予納台帳番号】 006194

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

 【物件名】 要約書 1

 【包括委任状番号】 9722147

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 文章分類装置および方法

【特許請求の範囲】

【請求項 1】 文章集合に含まれる複数の文章を、1つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置において、

前記各文章と前記各タームとの関係を 2 次元表現した D T マトリクスを生成する D T マトリクス生成手段と、

グラフ理論で用いられる D M 分解法に基づいて、前記 D T マトリクス生成手段で得られた D T マトリクスを変形することにより、変形 D T マトリクスを生成する D T マトリクス変形手段と、

この D T マトリクス変形手段で得られた変形 D T マトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類手段とを備えることを特徴とする文章分類装置。

【請求項 2】 請求項 1 において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するラベル生成手段をさらに備えることを特徴とする文章分類装置。

【請求項 3】 請求項 1 において、

前記変形 D T マトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力する文章編成手段をさらに備えることを特徴とする文章分類装置。

【請求項 4】 請求項 1 において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力する要約作成手段をさらに備えることを特徴とする文章分類装置。

【請求項 5】 請求項 1 において、

前記タームリストに対して任意のタームを追加または削除するタームリスト編集手段と、

このタームリスト編集手段による編集前後のタームリストを用いて前記 D T マ

トリクス生成手段によりそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力する指標生成手段とをさらに備えることを特徴とする文章分類装置。

【請求項6】 文章集合に含まれる複数の文章を、1つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置で用いられる文章分類方法において、

前記各文章と前記各タームとの関係を2次元表現したDTマトリクスを生成するステップと、

グラフ理論で用いられるDM分解法に基づいて前記DTマトリクスを変形することにより、変形DTマトリクスを生成するステップと、

前記変形DTマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章を同一分類として出力するステップとを備えることを特徴とする文章分類方法。

【請求項7】 請求項6において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項8】 請求項6において、

前記変形DTマトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項9】 請求項6において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項10】 請求項6において、

前記タームリストに対して任意のタームを追加または削除するステップと、

編集前後のタームリストを用いてそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力するステップとを

さらに備えることを特徴とする文章分類方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、文章分類装置および方法に関し、特に文章の内容に応じて各文章を分類する文章分類装置および方法に関するものである。

【0002】

【従来の技術】

高度情報化社会では、情報処理技術や情報通信技術の発展に伴い、電子化された膨大な量の情報を容易に入手できる環境が提供されつつある。このような環境を利用して入手した情報は、そのデータ量も膨大となるため、所望する情報を効率よくかつ正確に把握する必要がある。

情報の内容を解析する技術として、各情報を構成する文章の内容に応じて各文章を分類する技術が研究されている。

【0003】

従来、文章を分類する技術として、予め各分類の内容を示すラベルを用意し、各文章の内容を所定のアルゴリズムで解析し、用意した各ラベルごとにそれぞれの文章を分類するものが提案されている（例えば、非特許文献1など参照）。

これら技術は、文章の分類に際し、予め各分類の内容を示すラベルを用意し、各種の学習アルゴリズムを用いて、これらラベルを各文章に対して精度よく割り当てることにより、各文章をラベルごとに分類しようとするものである。

【0004】

なお、出願人は、本明細書に記載した先行技術文献情報で特定される先行技術文献以外には、本発明に関連する先行技術文献を出願時までに発見するに至らなかった。

【0005】

【非特許文献1】

永田昌明他,「テキスト分類－学習論理の見本市－」,情報処理,42巻1号,2001年1月

【非特許文献 2】

北研二他,「情報検索アルゴリズム」,共立出版,2002年

【0006】**【発明が解決しようとする課題】**

しかしながら、このような従来の文章分類技術では、予めラベルを用意する必要があるため、分類対象となる各文章の内容をある程度把握して適切なラベルを選択して設定しておく必要がある。したがって、このラベル選択に際し、文章量が多くその内容が広範囲にわたる場合には大きな作業負担を要するという問題点があった。また、分類に用いるラベルは主観的に選択されることから、得られる分類そのものが限定的となり、想定しうる範囲を超えた新たな観点から文章を分類できないという問題点があった。

本発明はこのような課題を解決するためのものであり、比較的少ない作業負担で、主観にとらわれることなく柔軟に分類できる文章分類装置および方法を提供することを目的としている。

【0007】**【課題を解決するための手段】**

このような目的を達成するために、本発明にかかる文章分類装置は、文章集合に含まれる複数の文章を、1つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置において、各文章と各タームとの関係を2次元表現したDTマトリクスを生成するDTマトリクス生成手段と、グラフ理論で用いられるDM分解法に基づいて、DTマトリクス生成手段で得られたDTマトリクスを変形することにより、変形DTマトリクスを生成するDTマトリクス変形手段と、このDTマトリクス変形手段で得られた変形DTマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章をそれぞれ同一分類として出力する文章分類手段とを備えるものである。

【0008】

ラベル生成手段をさらに備え、任意のクラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するようにしてもよい。

文章編成手段をさらに備え、変形DTマトリクスでの文章の並び順序に応じて、任意のクラスタに属する文章またはすべての文章を順に出力するようにしてもよい。

要約作成手段をさらに備え、任意の文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するようにしてもよい。

【0009】

タームリストの編集に際し、タームリストに対して任意のタームを追加または削除するタームリスト編集手段を設け、指標生成手段で、このタームリスト編集手段による編集前後のタームリストを用いてDTマトリクス生成手段によりそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力するようにしてもよい。

【0010】

また、本発明にかかる文章分類方法は、文章集合に含まれる複数の文章を、1つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置で用いられる文章分類方法において、各文章と各タームとの関係を2次元表現したDTマトリクスを生成するステップと、グラフ理論で用いられるDM分解法に基づいてDTマトリクスを変形することにより、変形DTマトリクスを生成するステップと、変形DTマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章をそれぞれ同一分類として出力するステップとを備えるものである。

【0011】

この際、任意のクラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するステップをさらに備えてもよい。

また、変形DTマトリクスでの文章の並び順序に応じて、任意のクラスタに属する文章またはすべての文章を順に出力するステップをさらに備えてもよい。

また、任意の文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するステップをさらに備えてもよい。

【0012】

タームリストの編集に際し、タームリストに対して任意のタームを追加または削除するステップと、編集前後のタームリストを用いてそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力するステップとをさらに備えてもよい。

【0013】

【発明の実施の形態】

次に、本発明の実施の形態について図面を参照して説明する。

[文章分類装置の構成]

まず、図1を参照して、本発明の一実施の形態にかかる文章分類装置について説明する。図1は本発明の一実施の形態にかかる文章分類装置の構成を示すブロック図である。

この文章分類装置1は、全体としてコンピュータからなり、制御部10、記憶部20、操作入力部30、画面表示部40、およびデータ入出力インターフェース部（以下、データ入出力I/F部という）50が設けられている。

【0014】

制御部10は、CPUなどのマイクロプロセッサとその周辺回路からなり、記憶部20に予め格納されているプログラムを実行して、上記ハードウェアとプログラムとを協働させることにより、文章分類処理のための各種機能手段を実現する。

記憶部20は、ハードディスクやメモリなどの記憶装置からなり、制御部10での処理に用いる各種情報を格納する。これら情報としては、分類対象となる各文章からなる文章集合21や、各文章の内容を把握するための複数の重要語からなるタームリスト22が記憶されている。

【0015】

操作入力部30は、キーボードやマウスなどの入力装置からなり、利用者の操作を検出して制御部10へ出力する。

画面表示部40は、CRTやLCDなどの画面表示装置からなり、制御部10での処理内容や処理結果を表示出力する。

データ入出力I/F部50は、外部装置（図示せず）や通信ネットワーク（図

示せず)と接続するための回路部であり、文書集合21やタームリスト22のほか、得られた処理結果や制御部10で実行するプログラムをやり取りする際に用いられる。

【0016】

制御部10には、機能手段として、DTマトリクス生成手段11、DTマトリクス変形手段12、文章分類手段13、ラベル生成手段14、文章編成手段15、要約作成手段16、タームリスト編集手段17、タームリスト生成手段18、および指標生成手段19が設けられている。

本実施の形態において、DTマトリクスとは、各文章Dと各タームTとの関係を2次元的に表現した行列を指す。

【0017】

本実施の形態では、上記関係は、文章D中におけるタームTの存在有無からなり、文章DとタームTとをそれぞれマトリクスの列と行に対応させ、ある文章D_iがあるタームT_jを含む場合には、DTマトリクスのj, i成分を「1」とし、含まない場合には「0」とすることにより、文章DとタームTの関係を表している。

さらに、このDTマトリクスを2部グラフの一表現形態と見なし、2部グラフのグラフ理論で用いられるDM分解法に基づきDTマトリクスを変形し、得られた変形DTマトリクス上に現れるクラスタに基づき、各文章Dを分類するようにしたものである。

【0018】

DTマトリクス生成手段11は、分類対象となる各文章D (Document) とタームリスト22を構成する各タームT (Term) とからDT (Document-Term) マトリクスを生成する機能手段である。

DTマトリクス変形手段12は、DTマトリクス生成手段11で生成されたDTマトリクスをDM (Dulmage-Mendelsohn) 分解法に基づき変形する機能手段である。

【0019】

DM分解法とは、具体的には、DTマトリクスに対し、行操作(行同士を入れ

替える操作) または列操作 (列同士を入れ替える操作) を施して、三角行列化する処理である。この三角行列化されたDTマトリクスを変形DTマトリクスと呼ぶ。

文章分類手段13は、DTマトリクス変形手段12で得られた変形DTマトリクス上に現れるブロック化されたクラスタに基づき、文章集合21の各文章を分類する機能手段である。

【0020】

ラベル生成手段14は、各クラスタごとに、当該クラスタに属する各文章Dと強連結の関係にあるタームTを、当該クラスタのラベルとして出力する機能手段である。

文章編成手段15は、変形DTマトリクスにおける文章Dの並び順に基づき、文章集合21の各文章を並び替えて出力する機能手段である。

要約作成手段16は、文章Dと強連結の関係にあるタームTを含む文を、当該文章Dの要約として出力する機能手段である。

【0021】

タームリスト編集手段17は、操作入力部30からの操作に応じて、記憶部20のタームリスト22に対するタームTの追加/削除を行う機能手段である。

タームリスト生成手段18は、記憶部20の文章集合21に含まれる各文章Dを解析して、各文章Dの特徴を効果的に表現する語すなわち重要語を抽出し、これら重要語からなるタームTを用いてタームリスト22を生成する機能手段である。

指標生成手段19は、タームリスト編集手段17で編集されたタームリストについて、その編集前後におけるDTマトリクスに基づき当該編集による分類への影響を示す指標を生成する機能手段である。

【0022】

[文章分類装置の動作]

次に、図2を参照して、本実施の形態にかかる文章分類装置の動作について説明する。図2は本実施の形態にかかる文章分類装置のDTマトリクス生成処理を示すフローチャートである。

制御部10は、操作入力部30からの指示に応じて、文章分類処理に用いるDTマトリクスを生成するため、図2のDTマトリクス生成処理を開始する。

まず、DTマトリクス生成手段11は、記憶部20に格納されている文章集合21を読み込むとともに（ステップ100）、タームリスト22を読み込む（ステップ101）。

【0023】

図3に文章集合21の構成例を示す。この例は、「ストレス」についてWeb上で多数の回答者に自由に文章を記述してもらったものを集計したものであり、各文章Dごとに当該文章Dを管理するための文章番号Diとその文章を記述した回答者の識別情報とが割り当てられている。

図4はタームリスト22の構成例である。このタームリスト22は、所定のアルゴリズムに基づき各文章Dを解析し、得られた重要語の種別とその前後関係とから各タームTを構成したものであり、各タームTごとに当該タームTを管理するターム番号Tjが割り当てられている。

【0024】

各タームTは、2つの重要語のうち、前方に位置するキーワード前と後方に位置するキーワード後からなり、それぞれのキーワードごとにそのキーワードの内容を示す単語とその単語の品詞属性種別とが規定されている。また、各タームTには、後述するタームリスト生成処理により文書集合21から算出された、文章分類に用いる上での重みを示す重要度が対応付けられている。

例えばターム「1」は、「ストレス」と「解消」という2つのキーワードからなり、その位置関係は「ストレス」が前方に位置するものと規定されている。

DTマトリクス生成手段11は、文章集合21内の各文章について、あるしきい値以上の重要度を持った重要度リスト22の各タームTが存在するか否かチェックし、その結果からDTマトリクスを生成する（ステップ102）。

【0025】

図5にDTマトリクスの構成例を示す。このDTマトリクス11Aは、行方向（縦方向）にタームTが並べられており、列方向（横方向）に文章Dが並べられている。そして、各文章DとタームTの交差位置に、当該文章Dにおけるターム

Tの存在有無が2進数で記載されている。ここでは、文章DにタームTが存在する場合は「1」が設定され、存在しない場合は「0」が設定されている。

したがって、この例によれば、例えば文章D1には、タームT4, T7が含まれていることがわかる。またタームT2は、文章D2, D4に含まれていることがわかる。

【0026】

続いて、DTマトリクス変形手段12は、このようにしてDTマトリクス生成手段11で生成されたDTマトリクス11Aを、DM分解法に基づき変形して変形DTマトリクス11Bを生成し（ステップ103）、これを記憶部20に格納して、一連のマトリクス生成処理を終了する。

一般に、グラフ理論では、2つの集合に属するそれぞれの点とこれら点を結ぶ辺とからなる2部グラフを、各点間の関連性に基づき分離する手法として、DM分解法が用いられる。

本実施の形態では、DTマトリクス11Aを、文章DからタームTへの辺により結びつけられた2部グラフの一表現形態と見なすことができることに着目し、グラフ理論におけるDM分解法をDTマトリクス11Aに適用し、得られた変形DTマトリクスに基づき文章Dを分類するようにしたものである。

【0027】

[DM分解処理]

ここで、図6および図7を参照して、2部グラフにおけるDM分解処理について説明する。図6はDM分解処理を示すフローチャートである。図7はDM分解処理の過程を示す2部グラフである。以下では、文章DおよびタームTからなる2つの点集合と、これら点を結ぶ辺とからなる2部グラフGを処理対象とし、これをDM分解法により複数のグラフに分離する場合を例として説明する。なお、これら処理では、制御部10内部のメモリまたは記憶部20から各種データを読み出して、制御部10で所定の演算を行い、その結果を再び記憶するという動作が繰り返し行われる。

【0028】

まず、図7(a)に示すように、処理対象となる2部グラフGの各辺について

、文章DからタームTへの有向辺を生成する（ステップ200）。そして、図7（b）に示すように、文章D側に点sを用意し、点sから文章Dの各点に対して有向辺を生成する（ステップ201）。同様にして、タームT側に点tを用意し、タームTの各点から点tに対して有向辺を生成する（ステップ202）。

【0029】

次に、これら辺を介して点sから点tへ向かう経路を検索する（ステップ203）。例えば図7（b）では、辺250, 251, 252からなる経路を介して点sから点tへ向かうことができる。このような経路が存在する場合は（ステップ203: YES）、当該経路を構成する各辺を削除するとともに（ステップ204）、当該経路上の文章DからタームTへの有向辺とは逆向きの有向辺を、初期状態で空の2部グラフである最大マッチングMに生成し（ステップ205）、ステップ203へ戻って次の経路を検索する。図7（c）では、有向辺251に対応する逆向きの有向辺253が最大マッチングMに生成されている。

ステップ203において、すべての経路の検索が終了して新たな経路が検索されなかった場合（ステップ203: NO）、最大マッチングMが完成したことになる。

【0030】

このようにして、図7（d）に示すような最大マッチングMを完成させた後、最大マッチングMに属する各有向辺254を処理対象Gへ含める（ステップ206）。これにより、図7（e）に示すように、処理対象Gにおいて、最大マッチングMとして選択された辺255については、文章DからタームTへの有向辺とその逆方向の有向辺とから構成されることになる。

次に、タームTの各点のうち最大マッチングMに用いられなかった点すなわち自由点256を選択し（ステップ207）、処理対象Gの各辺を介して当該自由点256に到達可能な点の集合をクラスタ260とする（ステップ208）。

【0031】

同様にして、文章Dの各点のうち最大マッチングMに用いられなかった点すなわち自由点257を選択し（ステップ209）、処理対象Gの各辺を介して当該自由点257に到達可能な点の集合をクラスタ262とする（ステップ210）

。

そして、残りの文章DおよびタームTの各点のうち、双方向に到達可能な経路を有する点集合すなわち強連結をなす点集合をクラスタ261とし（ステップ211）、一連のDM分解処理を終了する。

このようにして、公知のDM分解法では、各クラスタが所定の順序で生成され、三角行列化された変形DTマトリクスが得られる。

【0032】

制御部10では、以上のようにして、図2のDTマトリクス生成処理を実行することにより、DTマトリクス生成手段11で文章集合21とタームリスト22とからDTマトリクス11Aを生成するとともに、DTマトリクス変形手段12でDTマトリクスに対して図6のDM分解処理を適用することにより、各文章Dがクラスタごとに分離された変形DTマトリクス11Bを生成する。

【0033】

図8にDTマトリクス11Aと変形DTマトリクス11Bの例を示す。ここでは、各文章D_iにおいてタームT_jが存在する場合、その交点にドットが配置されており、タームT_jが存在しない場合は空白となっている。図8（a）のDTマトリクス11Aでは、ドットがランダムに分布しているが、図8（b）の変形DTマトリクス11Bでは、ドットが断片的ではあるが斜め方向に連続して密集しており、この部分270にクラスタが並んでいることがわかる。また、DTマトリクス11Bでは、左下側にドットが存在せず、右上側にドットが多く存在しており、上三角行列化されていることがわかる。

【0034】

[文章分類処理]

文章分類装置1の制御部10では、文章集合21进行分类する場合、まず前述のDTマトリクス生成処理（図2参照）を実行した後、図9の文章分類処理を実行する。図9は文章分類処理を示すフローチャートである。

まず、文章分類手段13は、DTマトリクス変形手段12で生成した変形DTマトリクス11B上に現れた各クラスタを識別する（ステップ110）。この際、各クラスタについては、変形DTマトリクス11Bを生成した際に分離した部

分グラフに基づき識別してもよく、変形DTマトリクス11B上のデータ（ドット）の並びから識別してもよい。

【0035】

図10に文章分類処理の説明図を示す。この例では、変形DTマトリクス11B上にクラスタ60が存在している。このクラスタ60は、2部グラフで表現した場合の部分グラフ61をなしており、他の文章やタームと関連性が小さい。なお、クラスタ境界が明確な完全グラフをなす場合もある。変形DTマトリクス11Bでは、列方向（横方向）に文章Dが並んでおり、クラスタ60の列方向に並ぶ文章DすなわちD363, D155, D157, D5, D13, D8が、このクラスタ60に属する文章Dとなる。

文章分類手段13は、識別された各クラスタに属する各文章からなる部分集合62を1つの分類として、文章集合21から抽出して分類し（ステップ111）、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連の文章分類処理を終了する。

【0036】

このように、本実施の形態では、変形DTマトリクス11B上でブロック化されたクラスタごとに、当該クラスタに属する各文章を1つの分類として抽出出力するようにしたので、各分類に対応したラベルを予め用意することなく各文章を分類できる。

したがって、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要性がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。

【0037】

また、これらクラスタは、複数のタームを橋渡しとして関連付けられた複数の文章から構成されているため、同一タームを含む文章を1つの分類として抽出することができるだけでなく、これら文章内にほぼ共通して存在する他のタームについても、そのタームを含む文章を同一分類として抽出でき、内容に共通性や関連性を持つ文章を1つの分類として容易に抽出できる。

したがって、従来のように予め用意したラベルの有無のみに基づき文章を分類する場合と比較して、そのラベルに限定された主観的な分類ではなく、想定する範囲を超えた新たな観点から文章の内容や話題に沿って柔軟に分類を行うことができる。

【0038】

[ラベル生成処理]

文章分類装置1の制御部10では、文章分類手段13で分類された各文章の分類ごとにラベルを生成する場合、まず前述のDTマトリクス生成処理(図2参照)および文章分類処理(図9参照)を実行した後、図11のラベル生成処理を実行する。図11はラベル生成処理を示すフローチャートである。

まず、ラベル生成手段14は、ラベルを生成する対象となる分類すなわちクラスタに属する各文章Dについて、これら文章Dと強連結の関係にあるタームTを変形DTマトリクス11Bから選択する(ステップ120)。

【0039】

図12にラベル生成処理の説明図を示す。この例では、任意の分類に属する文章を示す部分集合62について、各文章Dと強連結の関係にあるタームT(63)がそれぞれ選択されている。なお、強連結とは、変形DTマトリクス11Bで各文章Dをクラスタごとに分類した際、その2部グラフにおいて、文章DとタームTとが互いに双方向の辺で結ばれたペアをいう。通常、これら強連結をなす文章DとタームTとは、当該クラスタにおいて対角線上に並ぶ。

次に、選択した各タームTの単語を当該分類のラベル64として出力し(ステップ121)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連のラベル生成処理を終了する。

【0040】

このように、本実施の形態では、対象となる分類のクラスタに属する各文章と強連結の関係にあるタームTを、当該分類のラベルとして出力するようにしたので、本実施の形態のように予め用意されたラベルに基づき文章を分類するものではない場合でも、各分類の特徴を単語で表現した適切なラベルを容易に生成できる。

【0041】

[文章編成処理]

文章分類装置1の制御部10では、各文章Dの並びを編成する場合、まず前述のDTマトリクス生成処理(図2参照)を実行した後、図13の文章編成処理を実行する。図13は文章編成処理を示すフローチャートである。

まず、文章編成手段15は、変形DTマトリクス11B上での並びに基づき、各文章Dを並び替える(ステップ130)。

【0042】

図14に文章編成処理の説明図を示す。前述したように、DTマトリクスをDM分解法により変形して得られた変形DTマトリクス11Bにおいて、各文章DはタームTを仲立ちとして互いに関連性の高いものが隣接して並んでいる。

文章編成手段15は、このような変形DTマトリクス11Bに基づき並び変えられた文章Dを編成し、編成された各文章65を出力し(ステップ131)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連の文章編成処理を終了する。

【0043】

特に、変形DTマトリクス11Bには、文章DおよびタームTの並びに所定の半順序が存在する。DTマトリクス11Aは、タームTを変数とする文章Dの線形連立方程式を示す行列と見なすことができ、変形DTマトリクス11Bは、これら各方程式の解が求まる順序にほぼ沿った順序で文章Dが並び替えられた結果を示している。このことから、変形DTマトリクス11B上の文章Dの並びには、前後の文章Dとの関連性が高いことがわかる。

【0044】

このように、本実施の形態では、変形DTマトリクス上の文章Dの並びに基づき、各文章Dを並び替えて出力するようにしたので、共通のタームすなわち単語を持った関連性の高い文章が順に得られることになり、前後の文章Dと話題の共通性が得られる。したがって、内容が類似した文章が前後に並べられていることから、アトランダムに文章Dを読む場合と比較して、文脈が途切れることなく読むことができクラスタさらには文書集合全体の内容を容易に把握できる。

この際、任意のクラスツナワチ分類に含まれる各文章Dを文書編成の対象として1つの文書を生成してもよく、文章集合21に含まれるすべての文章Dを文章編成の対象として1つの文書を生成してもよい。

【0045】

[要約作成処理]

文章分類装置1の制御部10では、複数の文からなる任意の文章Dの要約を作成する場合、まず前述のDTマトリクス生成処理(図2参照)を実行した後、図15の要約作成処理を実行する。図15は要約作成処理を示すフローチャートである。

まず、要約作成手段16は、対象となる文章Dについて、前述したラベル生成処理と同様にして、その文章Dと強連結の関係にあるタームTを変形DTマトリクス11Bから選択する(ステップ140)。

【0046】

図16に要約作成処理の説明図を示す。通常、文章D(66)は、複数の文から構成されており、これら文のいずれかに文章Dと強連結のタームT(67)が含まれていることになる。この際、このタームTは文章Dの特徴を示していることになる。

要約作成手段16は、このタームTを含む文を当該文章Dから選択して、これら文を当該文章Dの要約68として出力し(ステップ141)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連の要約作成処理を終了する。

【0047】

このように、本実施の形態によれば、対象となる文章Dと強連結の関係にあるタームTに基づいて、そのタームを含む文を当該文章Dの要約として出力するようにしたので、文章Dの要約を極めて容易にかつ適切に作成できる。

【0048】

[タームリスト生成処理]

タームリスト生成手段18は、文章集合21からタームリスト22を自動生成するものである。

文章からその文章を特徴付ける重要語を抽出する方法として、各種のアルゴリズムが提案されている。例えば、各単語の重要度を算出し、その重要度に基づき重要語を選択する T F I D F (Term Frequency Inverse Document Frequency) などのアルゴリズムを用いてもよい。あるいは、言語学的な解釈に基づかないフレーズ (共起語) を、辞書を用いることなく抽出する K e y G r a p h というアルゴリズムを用いてもよい (例えば、非特許文献 2 など参照)。

【0049】

タームリスト生成手段 18 では、このような公知のアルゴリズムを用いてタームリスト 22 を生成する。本実施の形態では、これら単語を特定するため、各単語の品詞属性を形態素解析により予め求めておき、単語のとの品詞属性をペアとして重要語を構成している。また、本実施の形態では、2つの重要語の出現順序を規定したものをタームとして定義しており、これにより文章の内容をより適切にタームで表現可能となっている。

なお、このタームリスト 22 については、タームリスト編集手段 17 で、操作入力部 30 からの指示に基づき生成してもよく、データ入出力 I/F 部 50 を介して予め用意されたものを装置外部から入力するようにしてもよい。

【0050】

[指標生成処理]

タームリスト 22 は、変形 D T マトリクス 11 B を生成して文章を分類する上で重要なファクタとなることから、重要語リスト編集手段 17 で、このタームリストを編集可能としている。

本実施の形態では、編集されたタームリストについて、制御部 10 の指標生成手段 19 により客観的な評価値を算出し、その編集に対する指標を生成する。以下、図 17 を参照して、指標生成手段 19 における指標生成処理について説明する。図 17 は指標生成処理を示すフローチャートである。

【0051】

まず、タームリスト編集手段 17 により、タームリスト 22 についてターム T k を追加または削除し、新たなタームリストが生成されたものとする (ステップ 150)。指標生成手段 19 では、編集前後のタームリストのそれぞれについて

、DTマトリクス生成手段11によりDTマトリクスを生成し(ステップ151)、各DTマトリクスごとに平均文章類似度Qを算出する(ステップ152)。

平均文章類似度Qは、2つの文章 D_i 、 D_j 間の類似度 $\text{sim}(D_i, D_j)$ をすべての文章間について算出し平均したものであり、文章Dの数をNとした場合、Qは次の数1で算出される。

【0052】

【数1】

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N \text{sim}(D_i, D_j) \dots (1)$$

【0053】

この際、類似度 $\text{sim}(D_i, D_j)$ は、当該変形DTマトリクスに基づき、文章 D_i 、 D_j における各タームTの有無を0/1で示すベクトルをX、Yとした場合、例えば数2～数4により算出される。特に、数2はベクトルX、Yの内積を類似度とするもの、数3はベクトルX、YのDice係数を類似度とするもの、数4はベクトルX、YのJaccard係数を類似度とするものである。

【0054】

【数2】

$$\text{sim}(D_i, D_j) = |X \cap Y| = \sum_{i=1}^t x_i \cdot y_i \dots (2)$$

【0055】

【数3】

$$\text{sim}(D_i, D_j) = \frac{2|X \cap Y|}{|X| + |Y|} \dots (3)$$

【0056】

【数4】

$$\text{sim}(D_i, D_j) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \dots (4)$$

【0057】

このようにして、指標生成手段19は、編集前のタームリストから生成された

D Tマトリクスに基づき平均文章類似度 Q を算出するとともに、編集後のタームリストから生成されたD Tマトリクスに基づき平均文章類似度 Q_k を算出して、これらの差 $\Delta Q = Q_k - Q$ を算出し、指標値として画面表示部 40 から表示出力する（ステップ153）。

ここで、差 ΔQ がゼロより大きい場合は（ステップ154：YES）、編集後のタームリストから生成されたD Tマトリクスのほうが、各文章の類似度が大きくなり、各文章が効果的に分類できることから、当該編集は有効である旨を画面表示部 40 へ表示出力し（ステップ155）、一連の指標生成処理を終了する。

【0058】

また、ステップ154において、差 ΔQ がゼロ以下の場合は（ステップ154：NO）、編集後のタームリストから生成されたD Tマトリクスのほうが、各文章の類似度が小さくなり、各文章が効果的に分類できないことから、当該編集は無効である旨を画面表示部 40 へ表示出力し（ステップ156）、一連の指標生成処理を終了する。

なお、指標としては、 ΔQ だけを表示出力して作業者に編集の有効性を判断させるようにしてもよい。また当該編集に対する有効／無効だけを表示出力してもよい。

【0059】

このように、本実施の形態では、指標生成手段19により、編集前後のタームリストから生成されたD Tマトリクスに基づき平均文章類似度 Q を算出し、その変化により当該編集の有効性を示す指標を生成するようにしたので、タームリスト22に対する編集の有効性を容易に把握することができる。したがって、容易かつ適切にタームリストを編集でき、この編集により所望の意図や目的に応じて効率よく文章を分類することができる。

また、D Tマトリクスから得られた平均文章類似度に基づき指標を生成するようにしたので、文章を分類する必要がなくなり指標生成に要する処理を簡素化できる。したがって、当該編集に対する有効／無効を迅速に判断でき、タームリストの編集に要する作業負担を大幅に軽減できる。

【0060】

なお、平均文章類似度Qを用いて当該編集に対する有効／無効を判断する場合について説明したが、これに限定されるものではない。例えば文章を分類した結果、例えば分類数や1分類に属する文章数などにに基づき当該編集に対する有効／無効を判断するようにしてもよい。

【0061】

【発明の効果】

以上説明したように、本発明は、文章集合内の各文章とタームリスト内の各タームとからDTマトリクスを生成した後、そのDTマトリクスをDM分解し、得られた変形DTマトリクス上の各クラスごとに、当該クラスに属する各文章を1つの分類として抽出出力するようにしたので、各分類に対応したラベルを予め用意することなく各文章を分類できる。

したがって、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。

【0062】

また、これらクラスは、複数のタームを橋渡しとして関連付けられた複数の文章から構成されているため、同一タームを含む文章を1つの分類として抽出することができるだけでなく、これら文章内にほぼ共通して存在する他のタームについても、そのタームを含む文章を同一分類として抽出でき、内容に共通性や関連性のある文章を1つの分類として容易に抽出できる。

したがって、従来のように予め用意したラベルの有無のみに基づき文章を分類する場合と比較して、そのラベルに限定された主観的な分類ではなく、想定する範囲を超えた新たな観点から文章の内容や話題に沿って柔軟に分類を行うことができる。

【図面の簡単な説明】

【図1】 本発明の一実施の形態にかかる文章分類装置の構成を示すブロック図である。

【図2】 DTマトリクス生成処理を示すフローチャートである。

- 【図 3】 文章集合の構成例である。
- 【図 4】 タームリストの構成例である。
- 【図 5】 DTマトリクスの構成例である。
- 【図 6】 DM分解処理を示すフローチャートである。
- 【図 7】 DM分解処理の過程を示す 2 部グラフである。
- 【図 8】 DTマトリクスおよび変形DTマトリクスの例である。
- 【図 9】 文章分類処理を示すフローチャートである。
- 【図 10】 文章分類処理を示す説明図である。
- 【図 11】 ラベル生成処理を示すフローチャートである。
- 【図 12】 ラベル生成処理を示す説明図である。
- 【図 13】 文章編成処理を示すフローチャートである。
- 【図 14】 文章編成処理を示す説明図である。
- 【図 15】 要約作成処理を示すフローチャートである。
- 【図 16】 要約作成処理を示す説明図である。
- 【図 17】 指標生成処理を示すフローチャートである。

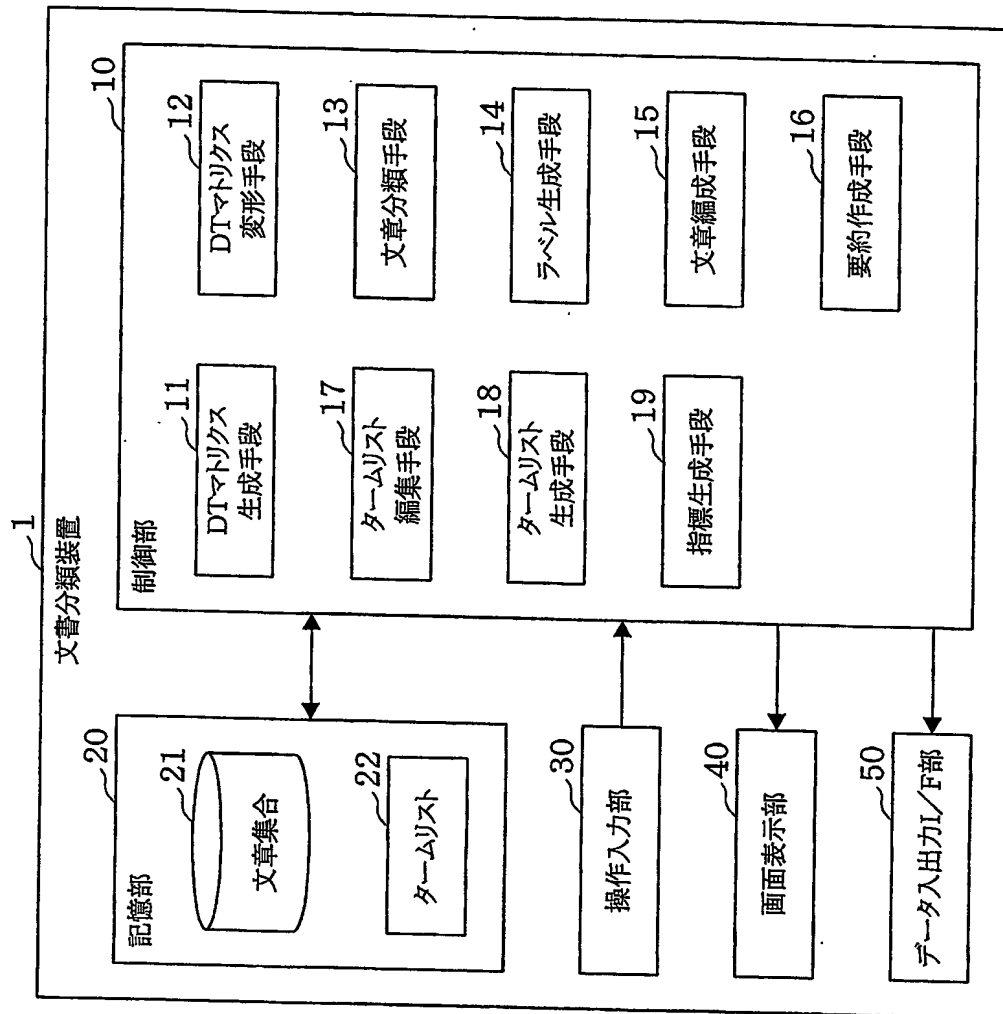
【符号の説明】

1…文章分類装置、10…制御部、11…DTマトリクス生成手段、11A…DTマトリクス、11B…変形DTマトリクス、12…DTマトリクス変形手段、13…文章分類手段、14…ラベル生成手段、15…文章編成手段、16…要約作成手段、17…タームリスト編集手段、18…タームリスト生成手段、19…指標生成手段、20…記憶部、21…文章集合、22…タームリスト、30…操作入力部、40…画面表示部、50…データ入出力 I/F 部、60…クラスタ、61…部分グラフ、62…部分集合（分類文章）、63…強連結ターム、64…ラベル、65…編成された文章、66…文章、67…強連結ターム、68…要約。

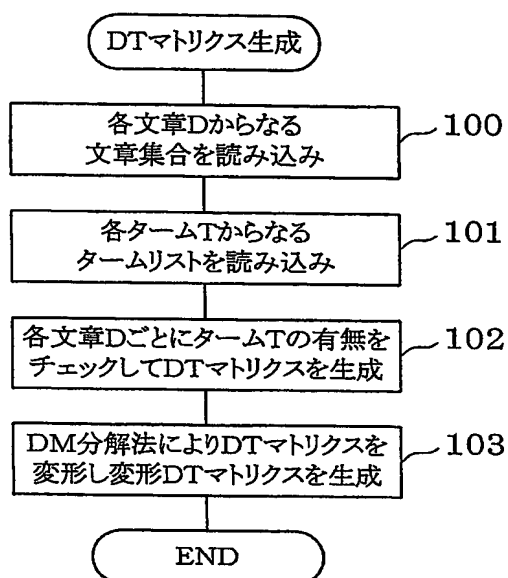
【書類名】

図面

【図1】



【図 2】



【図 3】

文章集合 21

DI	回答者	文章内容
1	W0039411	カラオケも楽しいですね。リズムにのって歌えるし間違っても誰も止めないですもんね。団でうたうのは周りの目も厳しいし頑張ら…
351	M0014787	現実逃避？なのかも知れませんが、私はストレスを感じたときは、ひたすら眠るようにしています。いい夢を見れば気分爽快なん…
289	M0010523	>好きな香りの入浴剤を入れて、のんびり入浴。>入浴しながら読書するとストレス解消になります。ぼくも本を持ち込んで読ん…
319	W0013732	マッサージって気持ちいいけど、ツボを心得ていない人にされるとかえってストレスになったりするのよね。マッサージも行きつけ…
57	W0039210	いいですね。気持ちよさそう!!! 私のように全然ばつとに当たらなければ、逆にストレスになってしまうかも… こういうストレスの…
72	W0039200	私もプールに通っています。ただがんばりすぎると逆に疲れるので三十分程度にして好きなように泳いだり歩いたりしています。…
111	W0012712	先日たった一泊ですが何年ぶりで旅行に行ってきました。食事の準備や布団の上げ下ろしもしなくていいのがこんな幸せな…
337	W0013147	>私のストレス解消法はなんといってもカーデニング！>旦那も子供も文句ばかり言うけど、花は何の文句も言わずに咲いてく…
360	W0015958	3歳になる息子がいます。公園に遊びに行ったりしますが、大の大人ならブランコなんか恥ずかしくて乗れませんが、子供と一緒に…
32	W0016759	私も不良主婦なんです。胸を張って家事は完璧ですなんて言えないのに、2~3ヶ月に1回くらいの割合で夜遊びします。7時ごろ…
:		

【図 4】

タームリスト 22

Tj	キーワード前		キーワード後		重要度
1	ストレス	名詞ー一般	解消	名詞ーサ変接続	256
2	解消	名詞ーサ変接続	ストレス	名詞ー一般	256
3	ストレス	名詞ー一般	仕事	名詞ーサ変接続	117
4	仕事	名詞ーサ変接続	ストレス	名詞ー一般	117
5	とき	名詞ー非自立ー副詞可能	ストレス	名詞ー一般	116
6	吸う	動詞ー自立	ストレス	名詞ー一般	99
7	的	名詞ー接尾ー形容動詞語幹	ストレス	名詞ー一般	88
8	行く	動詞ー自立	ストレス	名詞ー一般	86
9	寝酒	名詞ー一般	晩酌	名詞ー一般	85
10	晩酌	名詞ー一般	寝酒	名詞ー一般	85
11	人	名詞ー接尾ー助数詞	ストレス	名詞ー一般	83
12	子供	名詞ー一般	ストレス	名詞ー一般	77
⋮					

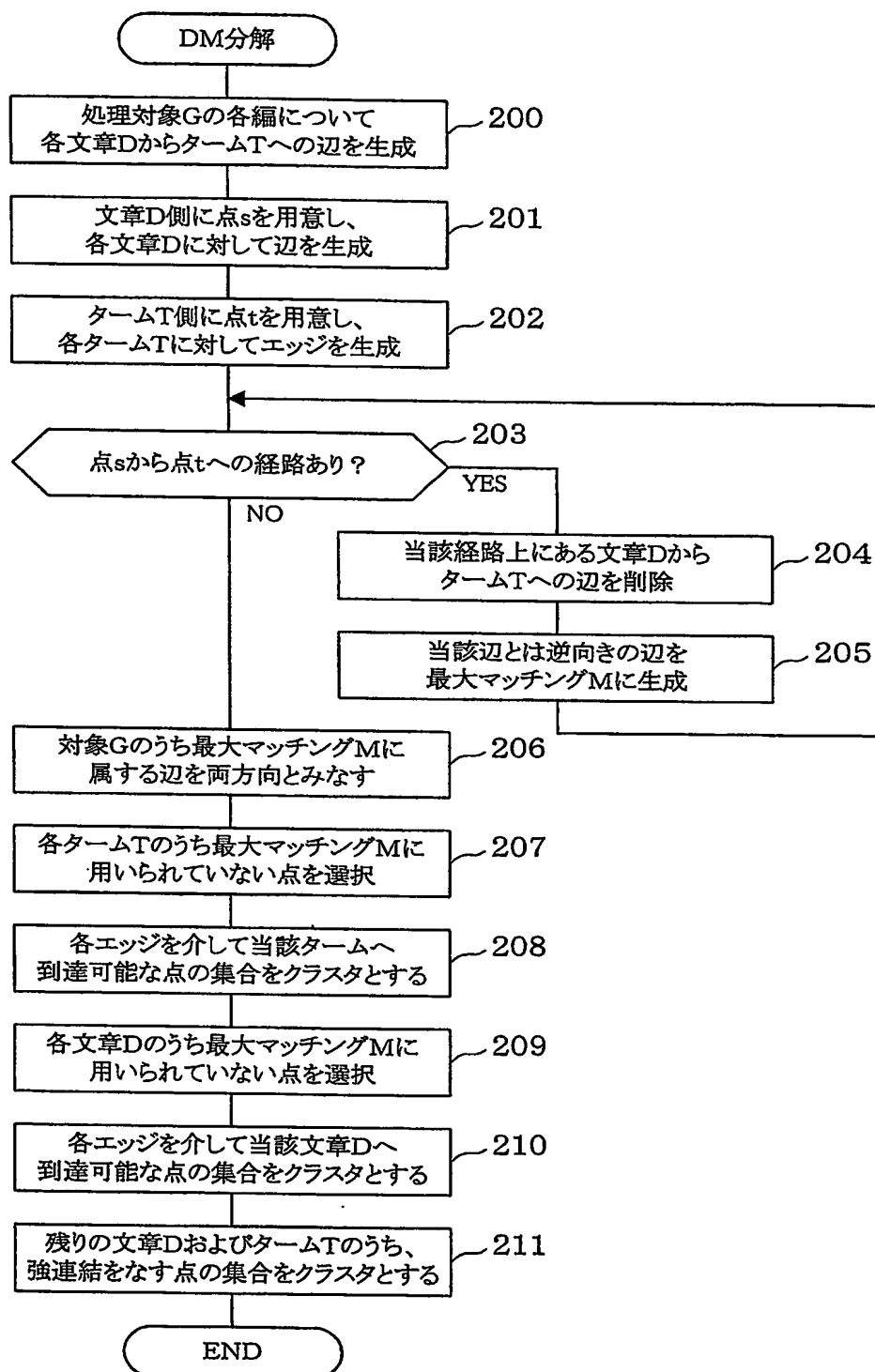
【図 5】

DTマトリクス 11A

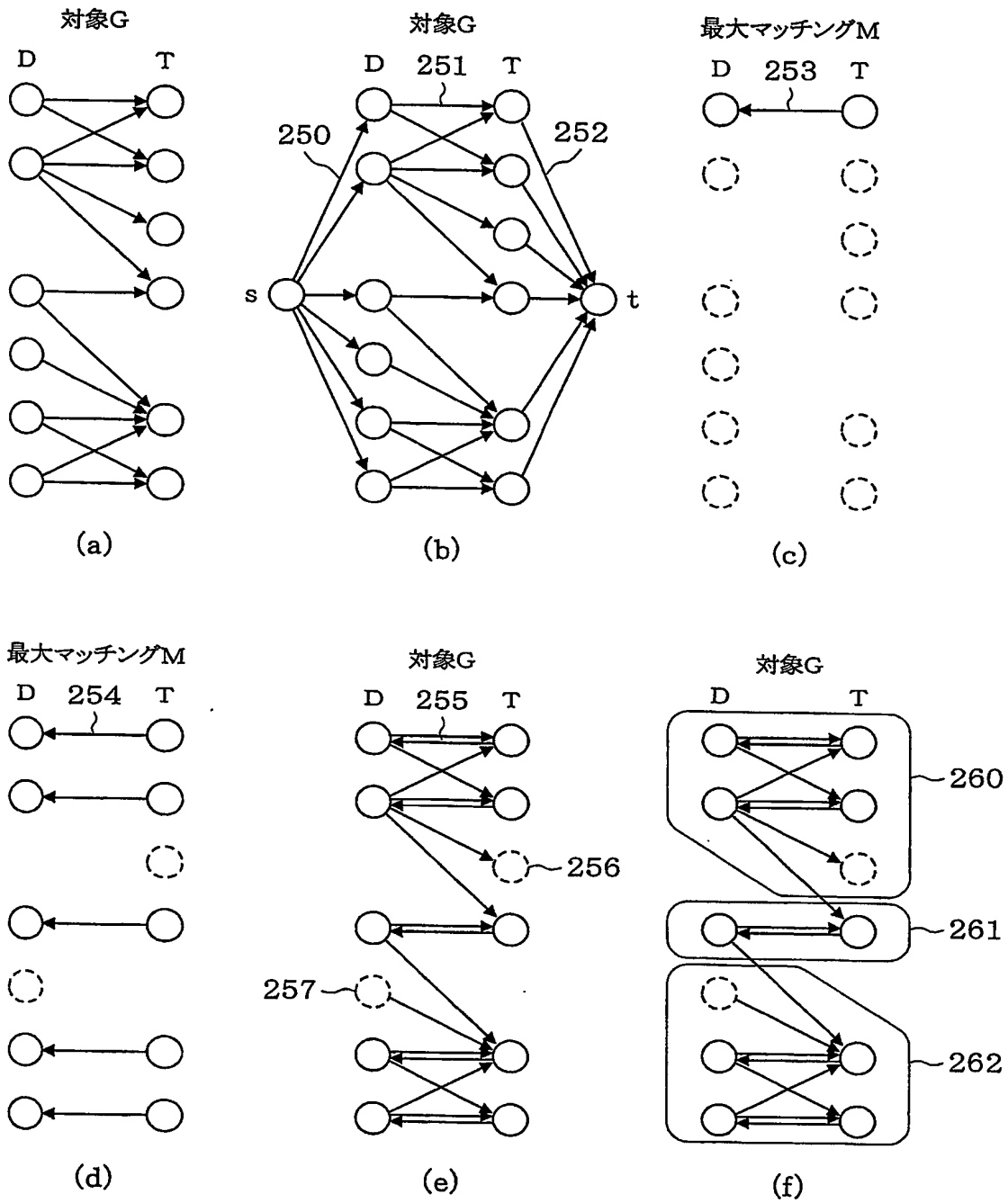
		文章D							
		D1	D2	D3	D4	D5	D6	-----	Dm
タームT	T1	0	1	0	0	0	0		0
	T2	0	1	0	1	0	0		0
	T3	0	0	0	0	0	1		0
	T4	1	0	0	0	0	0		0
	T5	0	0	0	0	0	0	-----	1
	T6	0	0	0	0	1	0		1
	T7	1	0	0	0	0	0		0
	T8	0	0	0	0	0	0		0
	⋮			⋮				⋮	⋮
	Tn	0	0	0	1	0	0	-----	0

0=文章DiにタームTjが存在しない
 1=文章DiにタームTjが存在する

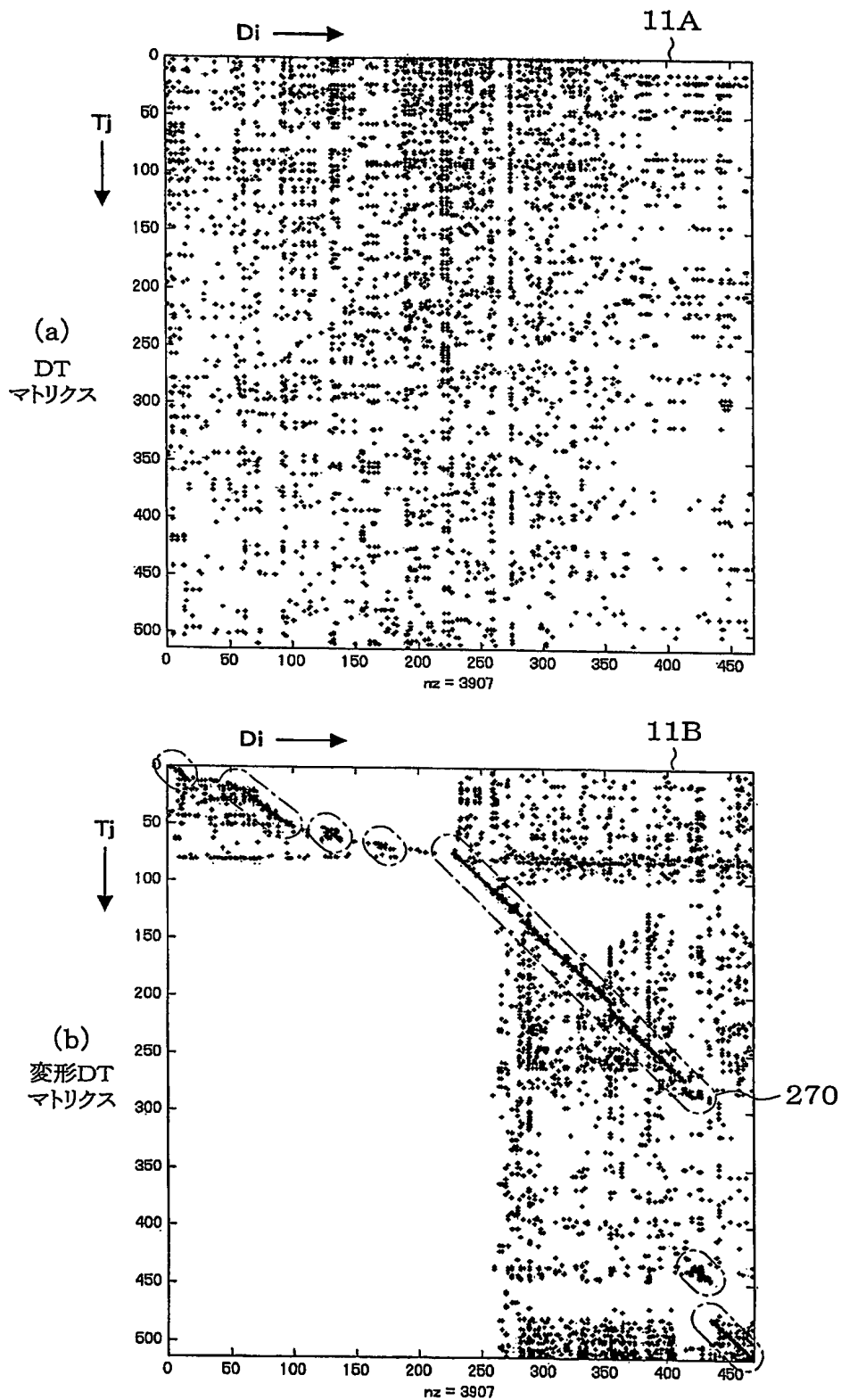
【図 6】



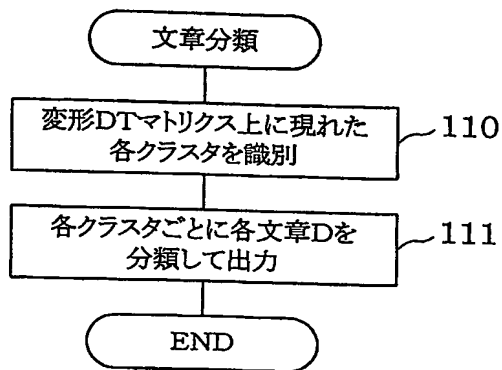
【図 7】



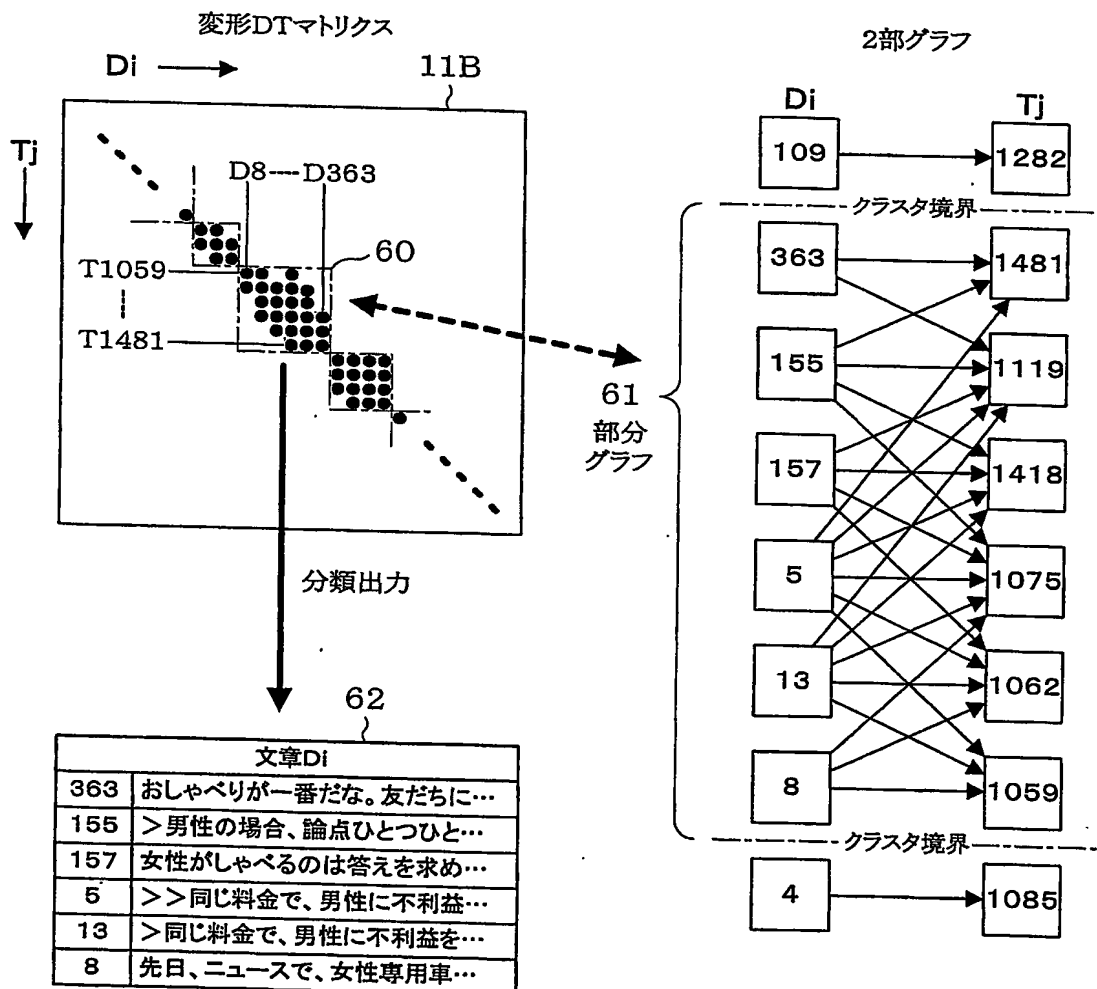
【図 8】



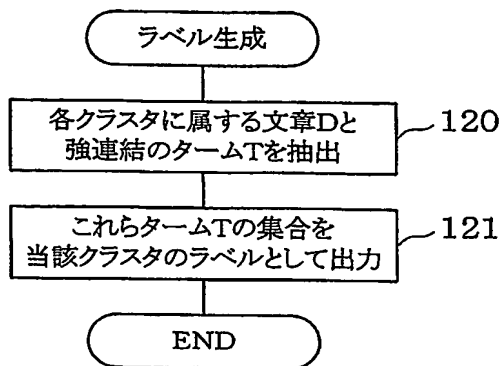
【図 9】



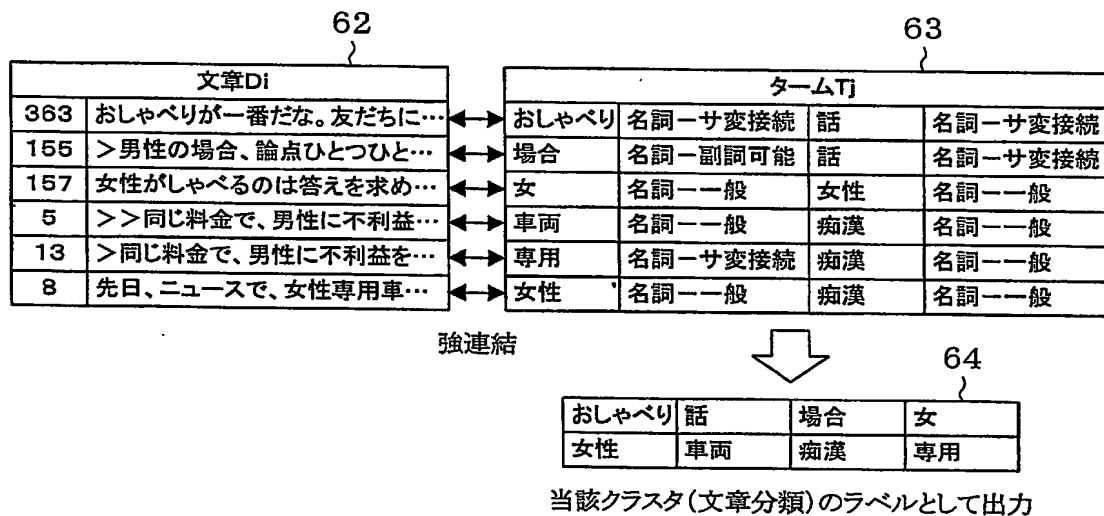
【図 10】



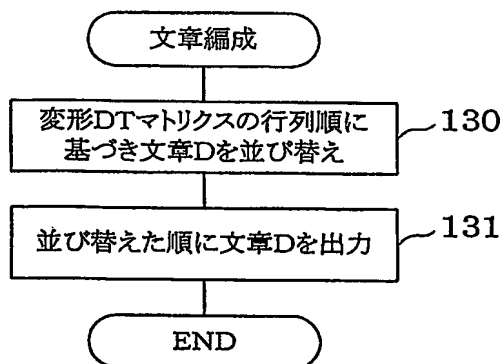
【図 1 1】



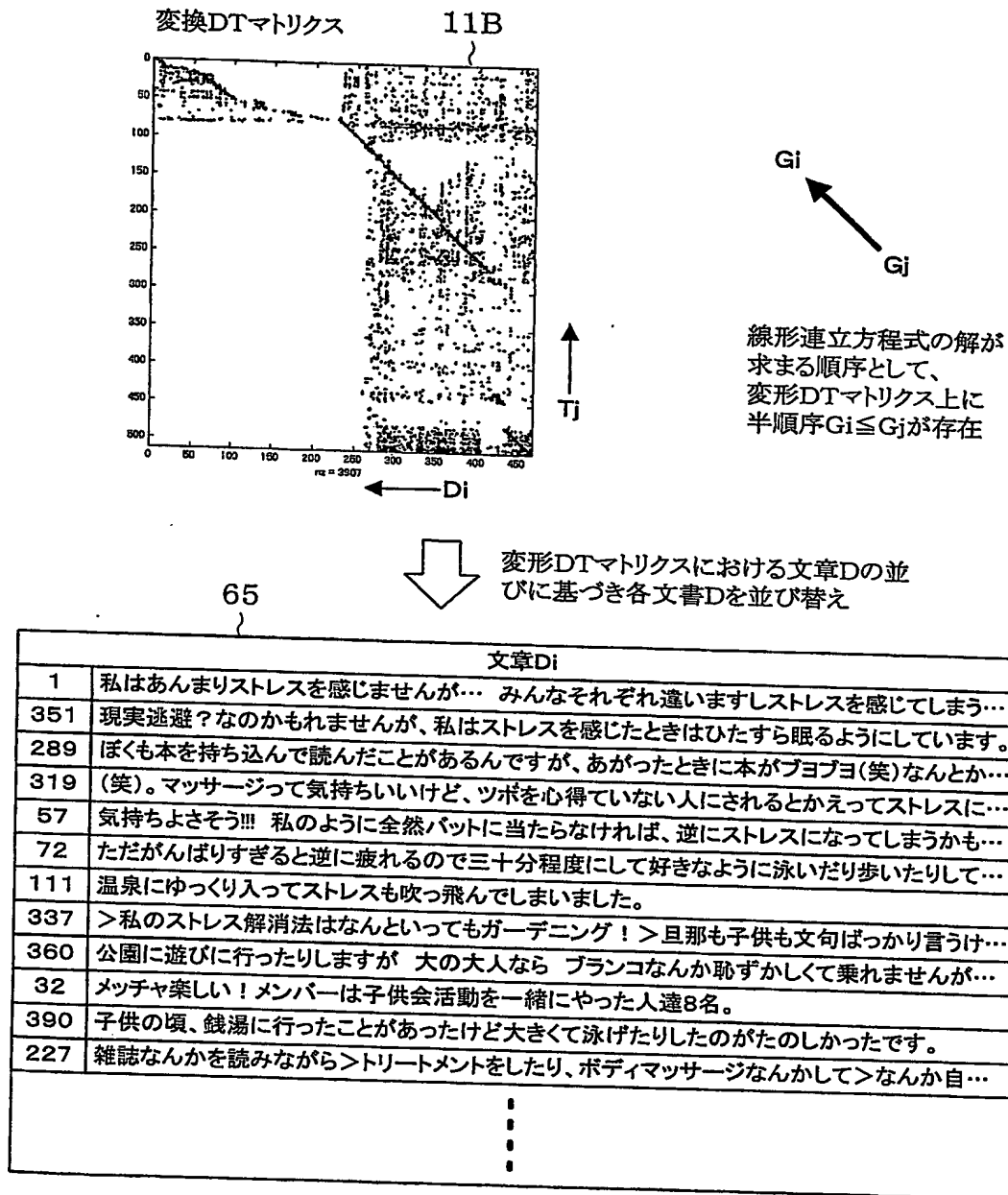
【図 1 2】



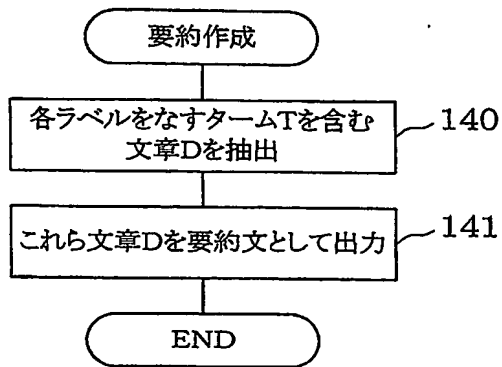
【図 1 3】



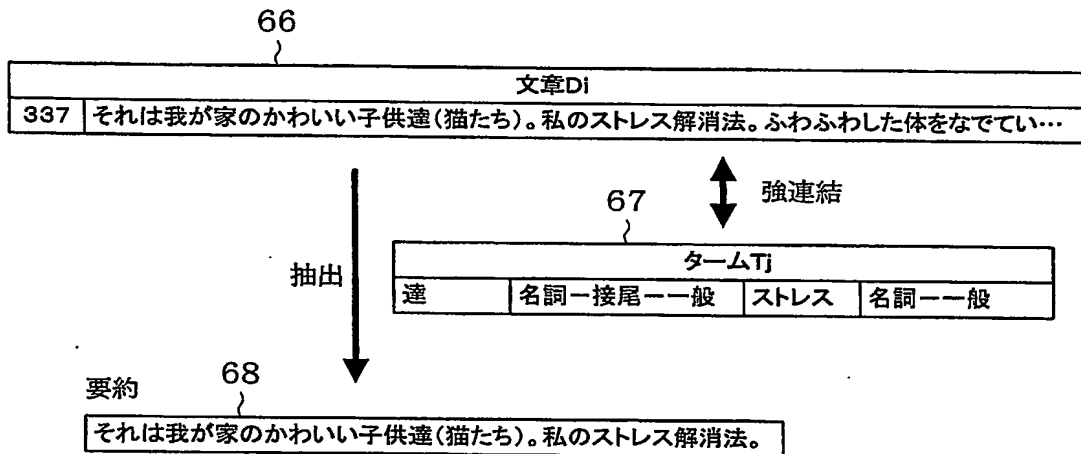
【図 14】



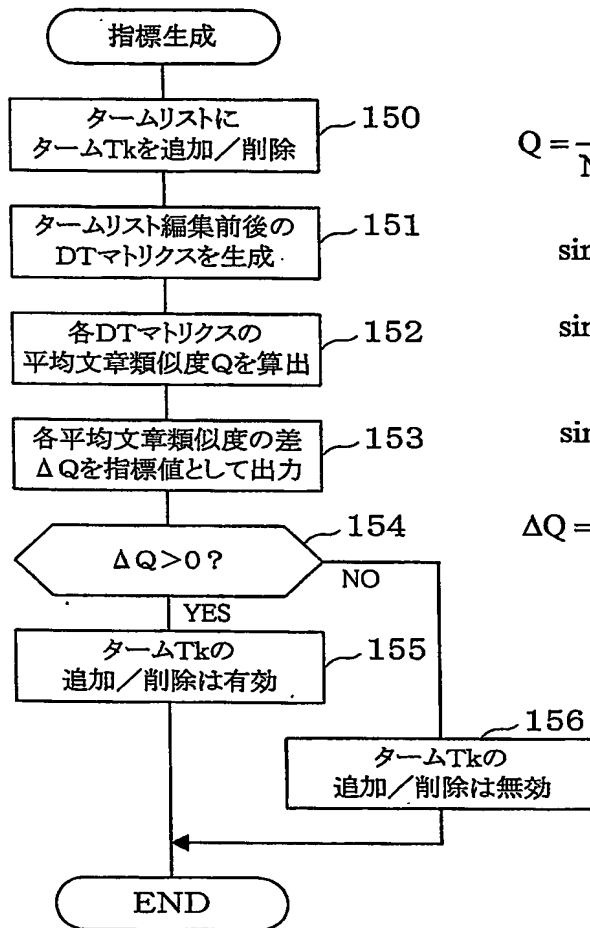
【図 15】



【図 16】



【図 17】



$$Q = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N \text{sim}(D_i, D_j) \dots (1)$$

$$\text{sim}(D_i, D_j) = |X \cap Y| = \sum_{i=1}^t x_i \cdot y_i \dots (2)$$

$$\text{sim}(D_i, D_j) = \frac{2|X \cap Y|}{|X| + |Y|} \dots (3)$$

$$\text{sim}(D_i, D_j) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \dots (4)$$

$$\Delta Q = Q_k - Q \dots (5)$$

【書類名】 要約書

【要約】

【課題】 比較的少ない作業負担で、主観にとらわれることなく柔軟に分類できるようにする。

【解決手段】 DTマトリクス生成手段 11 で、文章集合 21 内の各文章 D とタームリスト 22 内の各ターム T とから DTマトリクス 11A を生成し、DTマトリクス変形手段 12 で、その DTマトリクス 11A を DM 分解して変形 DTマトリクス 11B を生成する。そして文章分類手段 13 で、変形 DTマトリクス 11B 上に現れる各クラスごとに、当該クラスに属する各文章 D を 1 つの分類（部分集合）として抽出出力する。

【選択図】 図 1

特願 2003-189716

ページ: 1/E

出願人履歴情報

識別番号

[000006666]

1. 変更年月日

1998年 7月 1日

[変更理由]

名称変更

住所

東京都渋谷区渋谷2丁目12番19号

氏名

株式会社山武